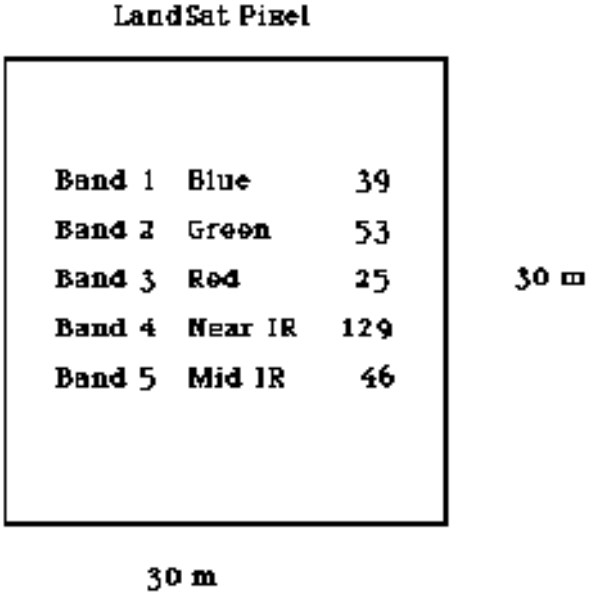
Unsupervised Classification "Clustering"

Each pixel in your LandSat TM image contains a wealth of information about the surface materials that reflected light from that pixel to the satellite sensors. Each pixel contains a value, from 0 to 255, for each TM band supplied with your image. If, for instance, your image contains data for five bands, then each pixel contains five pieces of data, each ranging from 0 to 255, as shown in the sample pixel diagram to the right.



This means that your image could contain 256^5 (that's approximately 1.1 billion) different possible spectral combinations. Each of these combinations does **not** represent a different type of land cover; most of these variations represent very small and, to us, "unseeable" differences in surface reflectance.

In most instances, your computer monitor will be displaying only 256 different colors, hence only 256 different pixels. Even set to "thousands" of colors, only a small part of the many different pixels can be displayed. Even if a monitor could display all the different possible pixels, your eyes simply could recognize only a small number of differences in their appearance.

Because there is a limited number of different land cover types (the Modified UNESCO Classifications scheme, MUC, contains about 130 different types), and no GLOBE study site will have all of those different land cover types, it is necessary to group pixels together into a smaller number of closely related "classes." This process, whereby pixels with similar spectral characteristics are grouped, is done in two different ways, by a supervised and unsupervised classification.

In a supervised classification, you "train" the software to recognize that certain types of pixels represent specific land cover types. This is done on the basis of your knowledge of your own area, and field work you may do. The software then classifies the pixels of your image into the groups you have specified. The MultiSpec tutorial provided with your GLOBE materials contains a section on supervised classification.

In an unsupervised classification, or "Clustering", we enter the number of groups, or "clusters," we wish to have, and certain other specifications. The software then examines the pixels in the image and groups them according to similar spectral characteristics. These groupings are not made on the basis of land cover, but on the similarity of the spectral characteristics of the pixels.

As part of your preparation of a land cover map for your 15 km x 15 km primary GLOBE study site, it is necessary for you to identify relatively large, homogeneous areas in your image for ground study and later use in a supervised classification. To do this, you will have MultiSpec cluster your image. This will help you locate areas to visit for ground verification studies.

Clustering

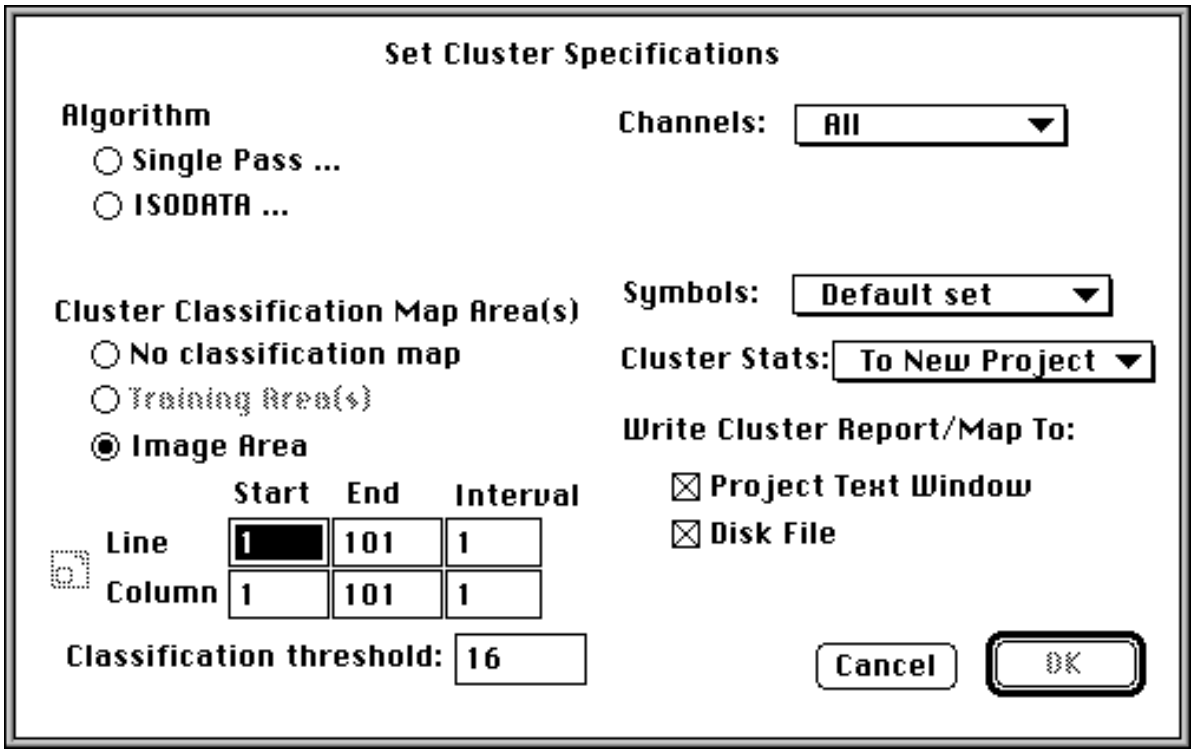
To demonstrate clustering, you will use a "sub-set" of the Beverly, Massachusetts image provided with your MultiSpec tutorial. This 101 x 101 pixel sub-image will allow the demonstration process to proceed more quickly than the clustering of a 512 x 512 image, and will allow you to follow exactly the steps outlined in this tutorial.

- Launch **MultiSpec** and **Open** the **beverlysubset.lan** image..
- Prom the **Project** menu, as shown to the right, select **New Project.**

Your clustering exercises are saved as projects and, when done, can be opened by **MultiSpec** as **Thematic Images**

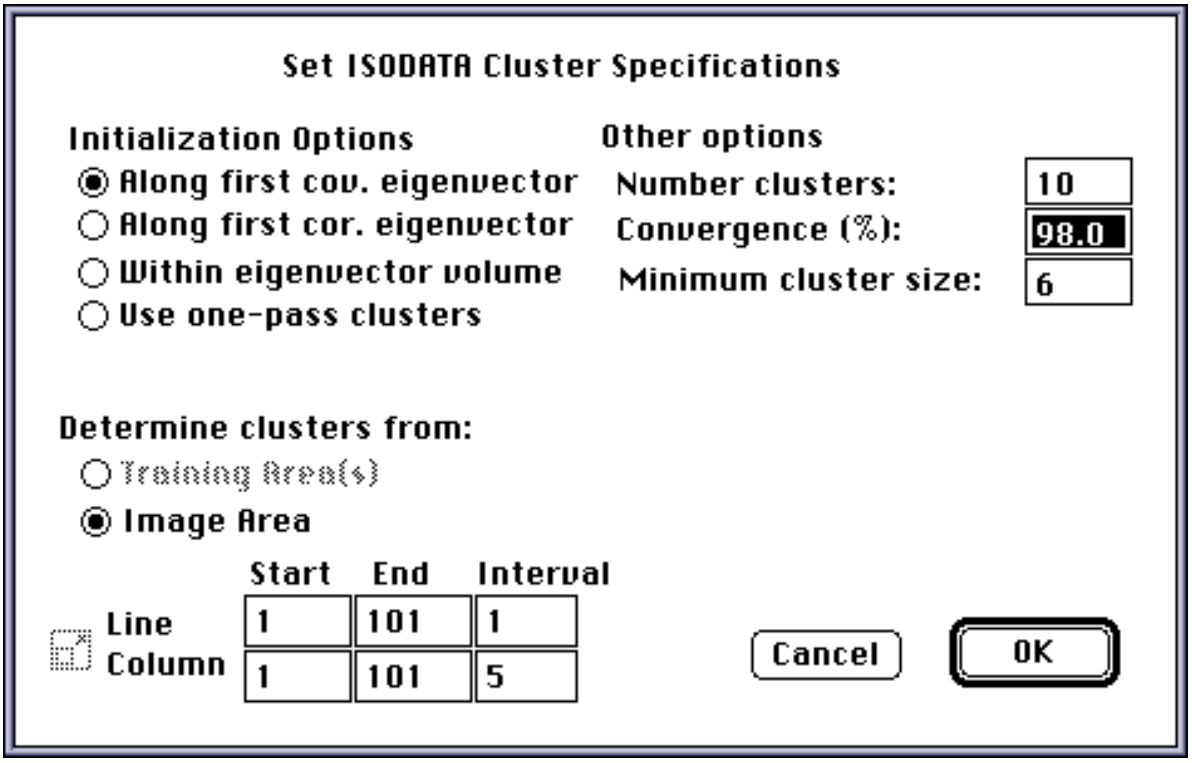
- From the **Processor** menu, select **Cluster**. "Clustering" is MultiSpec's terminology for an Unsupervised Classification. As shown on the next page, The **Set Cluster Specifications** window opens. It is in this window that you select a clustering "algorithm' (method by which the software clusters) and enter certain values for the software to use.

You must make certain settings in this window.



- First, be certain to click the **Image Area** button.
- Click to place an "X" in the **Disk File** box. This saves your project to disk.
- Be certain to select **To New Project** from the **Cluster Stats:** menu.
- Lastly, click the **ISODATA** button, as indicated by the cursor in the diagram above. **ISODATA** is the algorithm, or mathematical process, that MultiSpec will use in the clustering process.

A new window, the **Set ISODATA Cluster Specifications** window will open, as shown below.



It is in this window that you tell MultiSpec how you want the clustering to proceed. The information you need to provide is:

- Be certain that the **Image Area** radio button is checked, as shown above.
- Select "**Along first cov. eigenvector**." This is the *algorithm*[1] that MultiSpec will use in its clustering
- Leave the settings in the **Other options** boxes unchanged for this exercise.

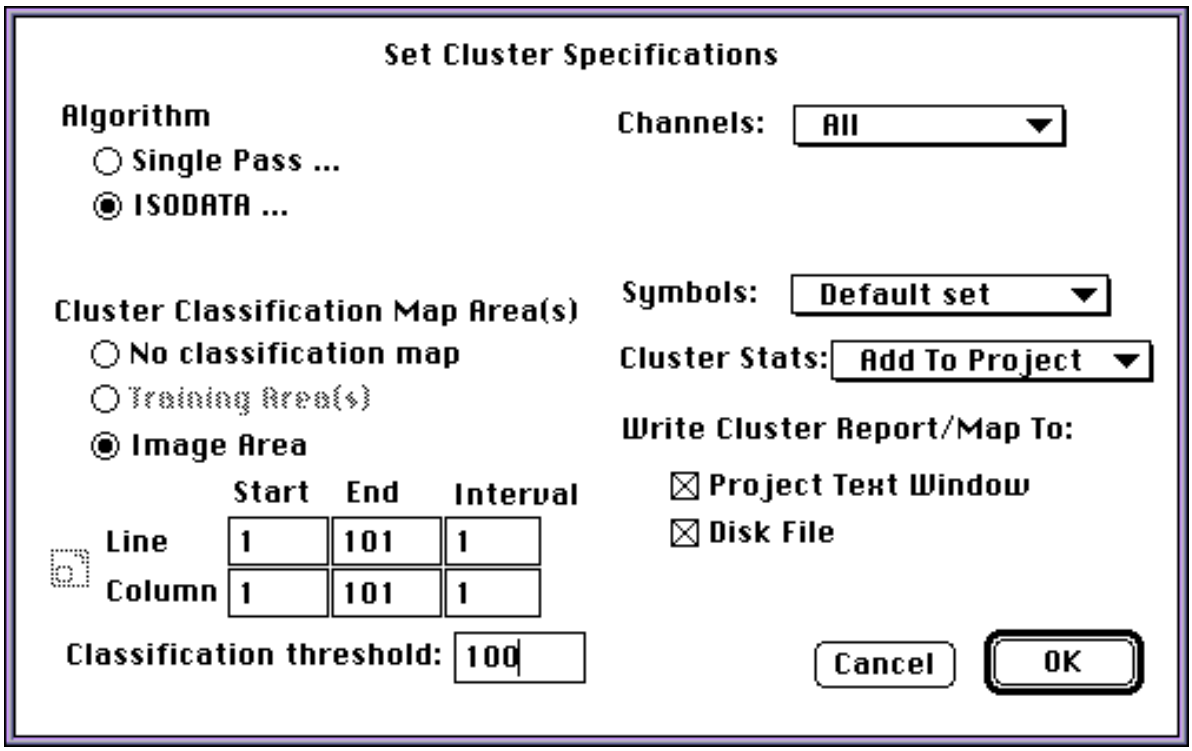
Notes: "**Number of clusters**" tells the software how many different groups you wish for the classification. The number 10 is used in this tutorial because we are classifying a small area. The number of clusters you will use when you cluster your 512 x 512 image will be discussed later.

During the classification, the program goes through the data over and over. This is called "iteration." Each iteration is called a "pass". The system makes "passes" through the image until a preset percentage of the pixels in the image are not changed during the pass. The clustering then ends. This percentage is called the "**Convergence**."

"**Minimum cluster size**" tells the system the smallest sized area to work with. Areas smaller than this minimum size will not be clustered.

[1] For a discussion of MultiSpec's algorithms, see "An Introduction to MultiSpec," by David Landgrebe and Larry Biehl, *Purdue Research Foundation*, 1995. This document may be downloaded from the Purdue/LARS WWW site at http://dynamo.ecn.purdue.edu/~Biehl/MultiSpec/

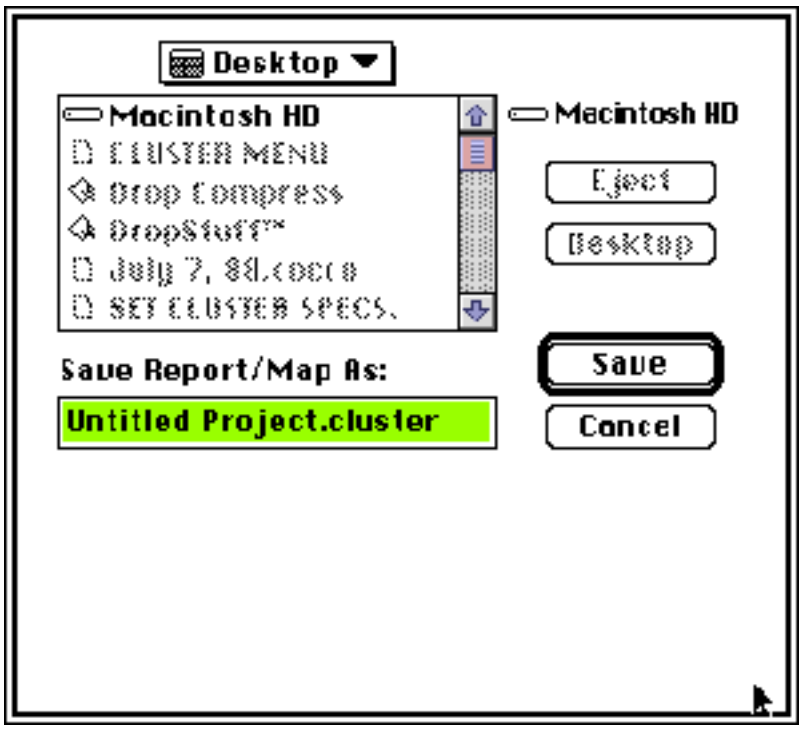
- After you have made these settings, click **OK.**
- The **Set Cluster Specifications** window appears *again*.



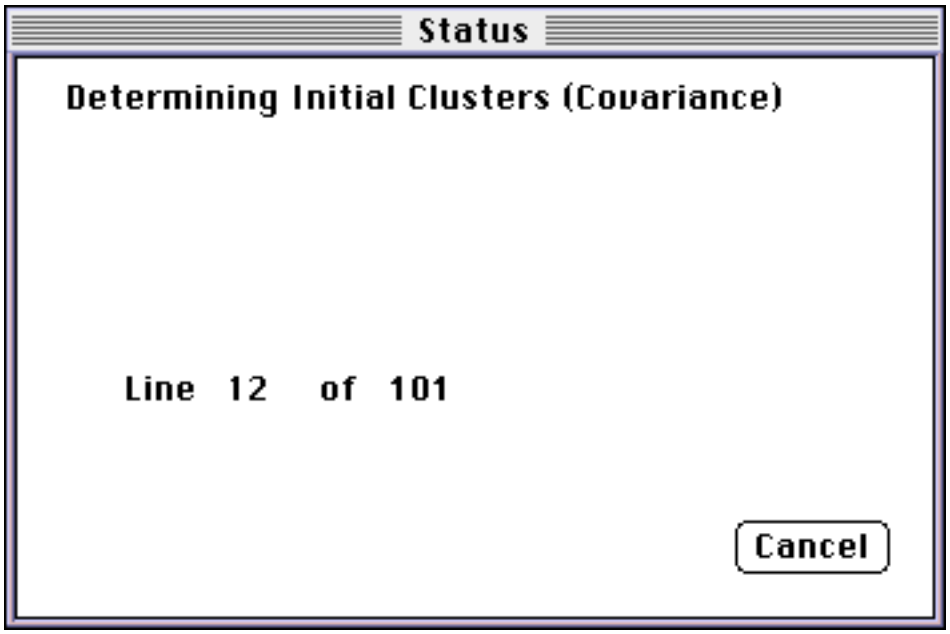
- In the lower left-hand corner of the box is the "**Classification threshold:** entry box. Change the value in this box to 100.

Setting this "threshold" value to 100 forces the system to assign every pixel in the image to one of the clusters. A value of less than 100 specifies the tolerance for assignment of pixels. A value of less than 100 will result in some pixels not being assigned to clusters. In this clustering, you are interested in large, fairly homogeneous areas, so individual pixels of slightly different spectral characteristics dotting the map are unnecessary.

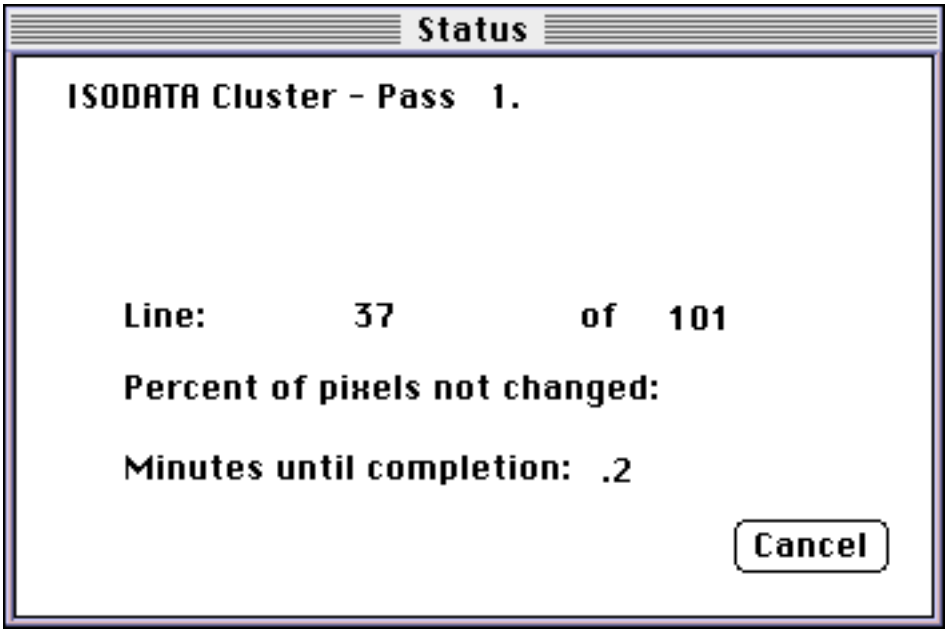
- The **Save File** dialog box appears, as shown below. There is a default name for your classified image file "**Untitled Project.Cluster**." You should change the Untitled Project" portion to something more descriptive, but leave the ".**Cluster**" extension to tell you what type of file this is.



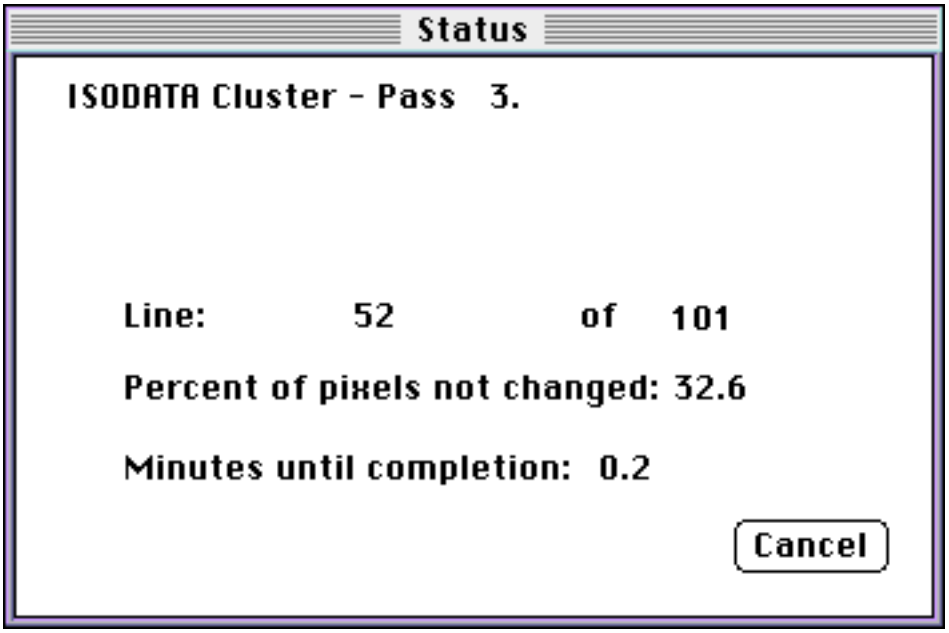
- The system then makes its first pass through the image to initially determine the clusters present, as shown in the **Status** box below.



- The first pass clustering **Status** box then appears, as shown. below. During the initial iteration, Pass 1, the "**Percent of pixels not changed**" shows no value. Also, a time is given for completion of the pass.



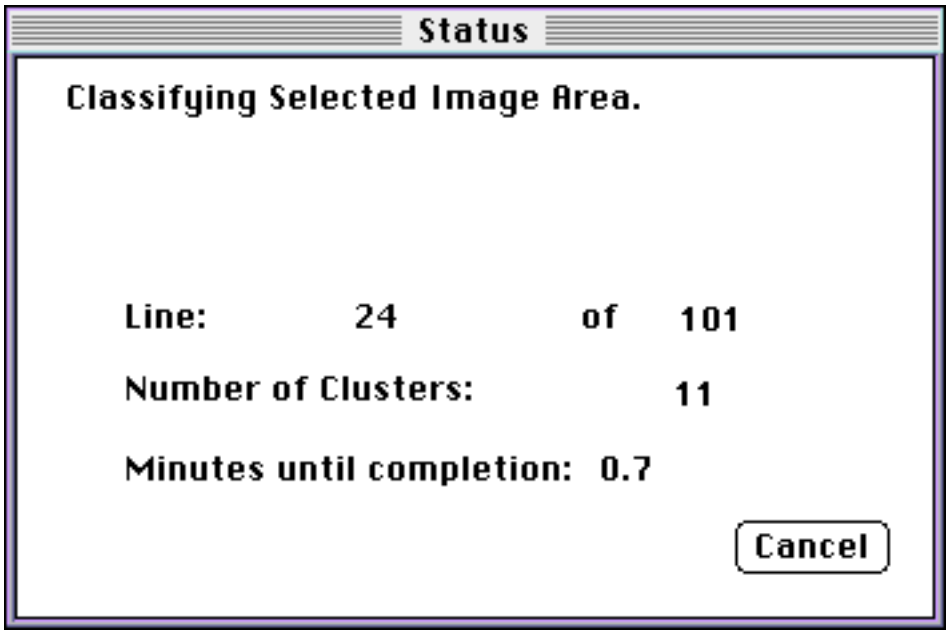
- The **Percent of pixels not changed**: entry does not change until the end of Pass 2. As the diagram below shows, a typical value of 30 to 40% will be achieved at this point.



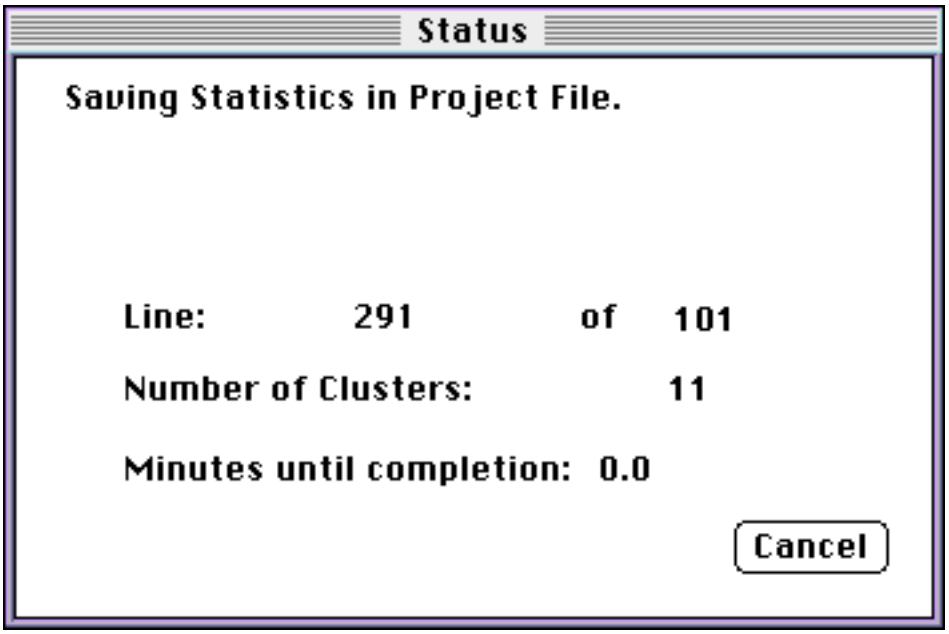
- During subsequent passes, the "Percentage of pixels not changed" value increases, until it reaches the value given in the "**Convergence (%)" specification**. The time for each pass to be completed is given in this window.

You can expect the system to make up to 12 to 14 passes to achieve a 98% Convergence. The time required for this process is dependent upon the processing speed of your computer. On a Power Mac 6100/66, running the "fat" version of MultiSpec (the version specially designed for Power Mac) with no other applications open, the process takes about 2 - 3 minutes. On a power book 150 (a very slow machine) the process can take several hours. If your machines are "old and slow" you should plan on the classification being the last exercise of the day. You can then let the processing go on overnight, and the results will be ready the next day.

- If you press "**Cancel**" during a pass, the "**Cancel**" button will darken, but you will see no immediate results. The clustering will be canceled **only when the current pass is completed**.
- After the clusters are determined, the system will display the "**Classifying Selected Image Area**" window, below. Here the system assigns individual image pixels to the clusters it has determined. Notice that it reports 11 clusters, when you specified 10. More on this later.



- After the clustering is complete, you will see the **Saving Statistics in Project File** window, shown below.



- The last message box you get will say "**Output text window being updated.**" The system then returns to your original image.

The Results of Clustering

There are two results of clustering:

A description of clustering activity and a "text map" in the **TEXT OUTPUT** window,

A clustered **Thematic** image.

- From the **Windows** menu, select **Text output**. Scroll to the top of the window, and you will have statistics describing the clustering and its results. A part of the text output for the sample clustering is shown below. In it are listed the number of clusters produced and the average value (mean) of the pixel values for each band in each of the classes.

Final cluster class statistics.

Cluster	Pixels	Channel Means				
		1	2	3	4	5
1	46	238.8	244.5	242.3	162.9	226.7
2	59	215.3	203.2	201.3	118.1	153.9
3	160	155.3	150.4	140.8	142.0	153.2
4	139	118.4	144.4	119.5	227.2	233.9
5	143	112.7	110.4	100.3	138.2	132.1
6	255	89.9	97.8	81.0	182.6	150.7
7	383	67.8	84.5	57.0	232.3	160.9
8	539	60.8	71.5	48.3	198.2	135.5
9	281	60.8	65.3	46.7	153.3	108.8
10	36	69.7	55.7	35.4	19.9	19.6

Number classes = 11

Note that 10 classes are listed, but the system says that 11 classes were used. The 11th class is reserved for "**Thresholded**" classes. These are areas that were not classified into any of the clusters produced by the clustering process. In this clustering, however, you set the threshold at 100, so that the "thresholded" cluster contains no pixels.

Also produced is a text map of the clustered area. The system assigns a number or letter to each of the clusters, and then displays a map of the clustered area using this code. For the clustered Beverly.sub image, the code is shown below.

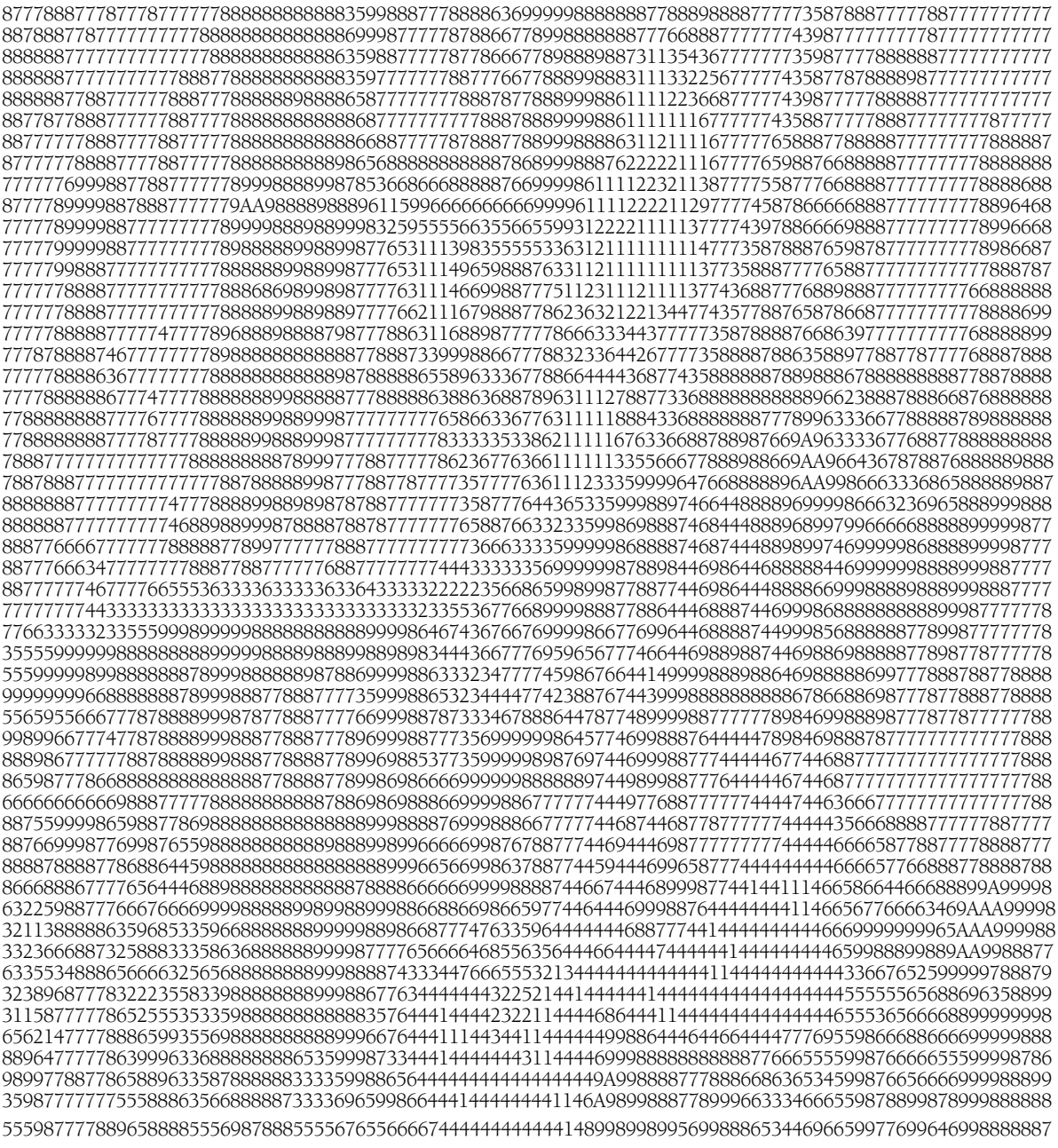
Classes used:

1: Cluster 1	1
2: Cluster 2	2
3: Cluster 3	3
4: Cluster 4	4
5: Cluster 5	5
6: Cluster 6	6
7: Cluster 7	7
8: Cluster 8	8
9: Cluster 9	9
10: Cluster 10	A
11: Thresholded	

A *portion* of the text map of the clustered area is shown below, in 9 point type. By holding the page at arm's length, you can see that it shows large areas of homogeneous land cover. This text map can be printed and hand colored to show you the location of areas to investigate. However, you will have to print your image in parts; a text map 512 characters wide and 512 lines long would be too large for most printers.

Classification of Selected Area

Lines 1 to 101 by 1. Columns 1 to 101 by 1.



Also produced is a classification summary, shown below, which gives the number of pixels in each cluster, and the number of pixels that were not classified.

Classification summary

Cluster 1 classification size: 247

Cluster 2 classification size: 252

Cluster 3 classification size: 770

Cluster 4 classification size: 676

Cluster 5 classification size: 789

Cluster 6 classification size: 1277

Cluster 7 classification size: 1907

Cluster 8 classification size: 2703

Cluster 9 classification size: 1366

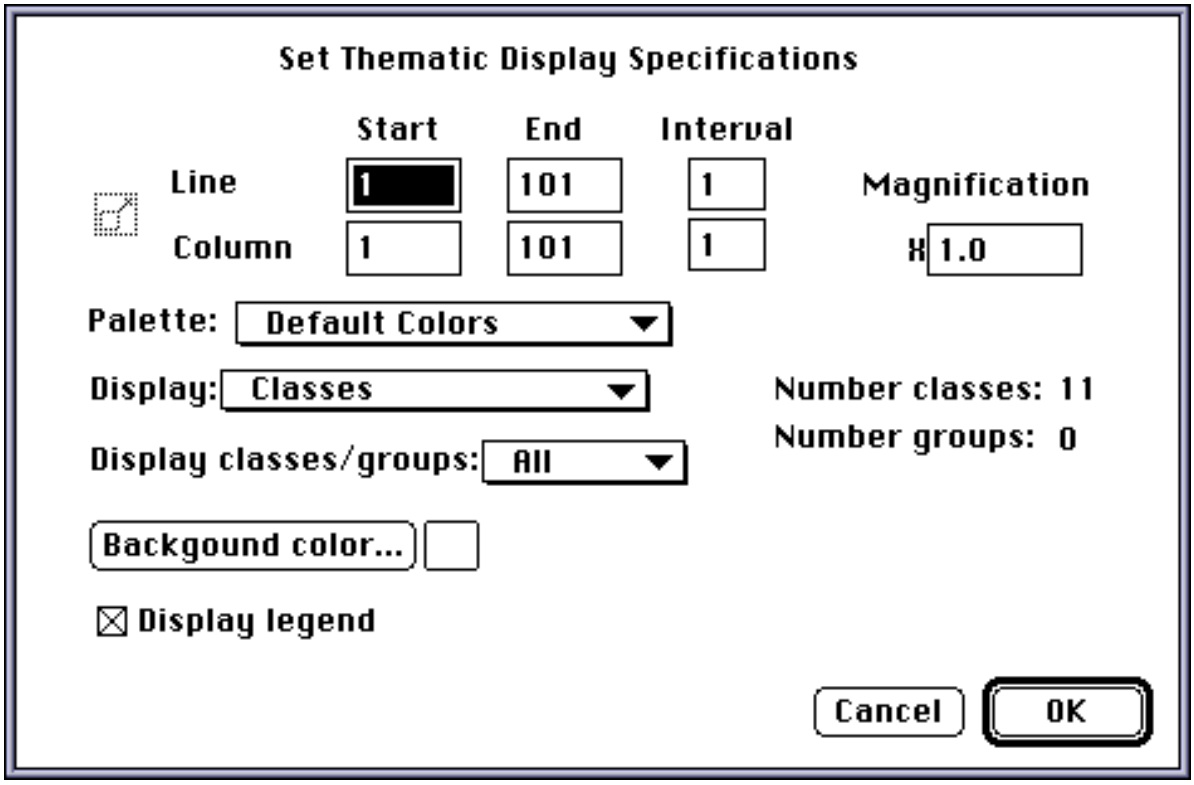
Cluster 10 classification size: 214

Number of pixels not classified = 0

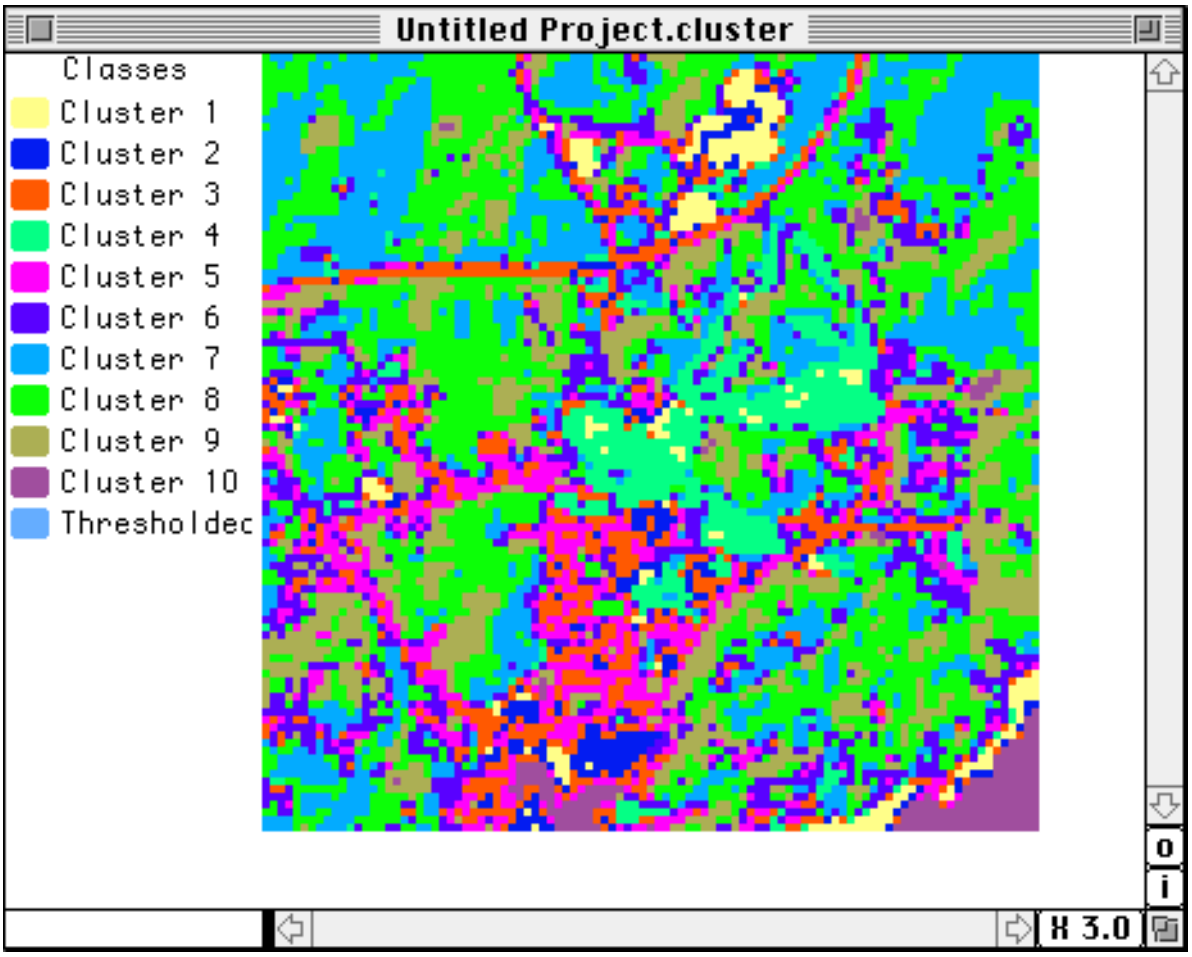
In this case, remember, there are 0 unclassified pixels, because you set the **Classification threshold** to 100.

Examining the Clustered Image

- From the File menu, select **Open Image**.
- Select the **.Cluster** file name you used earlier, and click **Open.**
- The **Set Thematic Display Specifications** window opens, as shown below. You can experiment later with some of the other palettes in this menu, but for now accept the default settings and press **OK**.



- Your clustered image opens, as shown below.



- Notice that there are 10 numbered classes, and the 11th "Thresholded" class. Each class is assigned a color by the system which has nothing whatsoever to do with what the cluster represents. The clusters are produced and arranged in order of descending level of brightness. That is, clusters near the top of the list represent surface materials that are "brighter' (have greater reflectance) than those near the bottom of the list.
- You may change any color by double-clicking on the color box in front of each cluster identification. You will get the standard Apple color selection window for your operating system. If you are not familiar with this color selection system, see your computer Users Guide.
- You may print the image from the **File** menu. When you do, the clustering key will be printed along with the image.
- You may use some of MultiSpec's regular tools with this Thematic Map. Such tools as: the **Zoom** feature, and **Show Selection Coordinates** function normally. The **New Selection Graph** feature will show a plot with only one piece of data. This map is no longer "multispectral." Each pixel no longer contains data for different LandSat bands, or channels. Each pixel contains only one value, which identifies its color.

- If you do a clustering with a larger number of classes, you may not be able to see them all in the "**Classes**" column. To scroll through this column:
  - Move your cursor into the column
  - Hold the mouse button down
  - Drag to either the top or bottom of the column.

The classes will scroll up and down.

  - It is sometimes difficult to tell which color in the **Classes** column matches colored areas in the image. To match classes to their image areas:
  - Place the cursor over any color box in the "Classes" column.
  - Hold down the shift key: The cursor changes to an "eye."
  - Depress the mouse button, and the areas in the image of that class will "blink," or turn to white.

- You and your students will probably want to prepare a thematic map from this clustered image in which you identify some of the clustered areas by their actual land cover. To do this, you may save the image as a **TIFF** file from the **File** menu. This process does not save the clustering key, only the image area will be saved. The TIFF file may then be brought into any one of a number of paint or draw programs to be "fancied up" as a thematic map.
- If you wish to have an image that contains the clustering key, and can also be moved into paint or draw programs you can capture the entire screen using Apple's "**Shift-Command-3**" feature. Hold down the Shift key, the Command key, and press "3." A "snapshot" of the screen is saved as a PICT file on your fixed drive and may be opened by any program that will handle PICT files. For PC computers, there are a number of programs that will accomplish the same screen "snapshot, and for Macintosh users there are several programs that do "screen captures" in a more flexible manner than the system's "Shift-Command-3" utility.

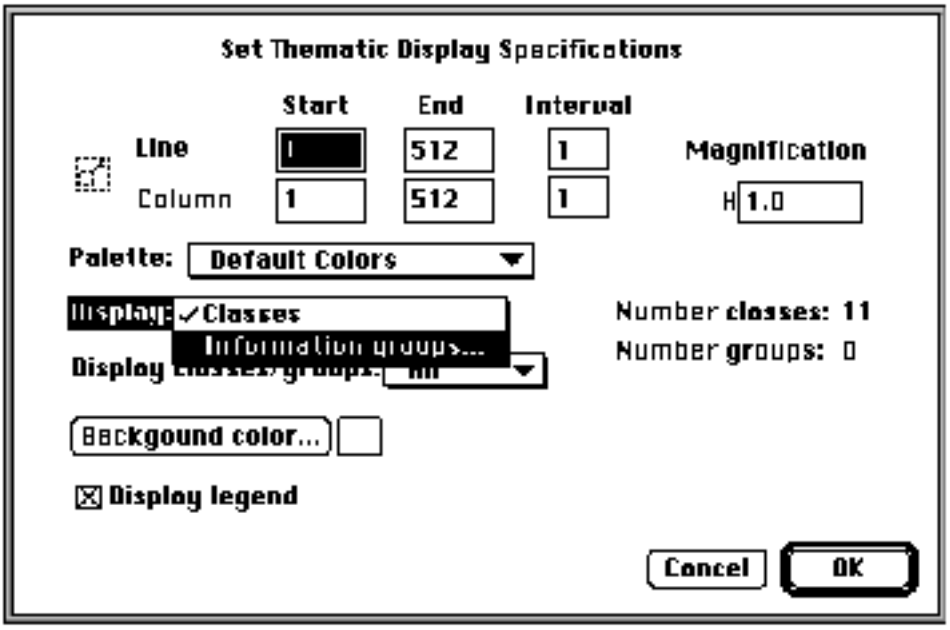
How Valid Is the Clustering Process?

It is necessary for you to be confident that this process of "unsupervised classification" actually yields clusters that are related to land cover types. To this end, included with this tutorial is a file named "**Beverly9subset.class**". This is the same image that you have clustered, only this image was prepared with a **supervised** classification by an individual very familiar with the land cover types in the area.

- In your clustered image, zoom in to 3X.

With your clustered image open, open the **Beverly9subset.class** image.

- From the **File** menu, select Open.
- Locate the **Beverly9subset.class** image, and **Open it.**
- When the **Set Thematic Display** window opens, as shown below, select **Information Groups** from the **Display** menu.



- Re-size and position each image so they are side-by-side on the screen.
- Compare the areas identified in the supervised classification to the clusters produced by the system in your unsupervised clustering.

You should see that the unsupervised clustering provides, at least in this case, a good indication of the locations of large areas of uniform land cover that could be investigated for verification studies.

How Many Clusters Do I Use?

Most regions the size of your 15 km x 15 km Primary GLOBE study site do not generally demonstrate a large number of different land covers. When you first perform a clustering on your 512 x 512 image, use the same values as you used in this tutorial. Examine the results in light of your knowledge of your own area. Do some field work and look at the areas your clustering suggests are fairly large and homogeneous. Compare your findings to the MUC classification scheme. Only if you feel that this clustering does not adequately represent the land covers in your area should you increase the number of clusters, and then 12 to 14 clusters should be sufficient to do the job.

Reporting the Data

In order to report your data, you must make some "sense" out of the clusters determined by this unsupervised process. You can then re-label the clusters as what type of land cover they represent. The process involves the following steps:

- Desk Verification
- Field Verification
- Renaming the Clusters
- Send in your completed map.

Desk Verification

This process involves your use of local maps (topographic, land cover, soil, political, etc.), other local references (aerial photos, people, agencies, etc.) and the combined experiences of both you and your students to identify some of the clusters produced by MultiSpec. Use whatever resources you can to identify them. Remember that your identifications should correspond to the level IV of the **MUC** (Modified Unesco Classification) scheme.

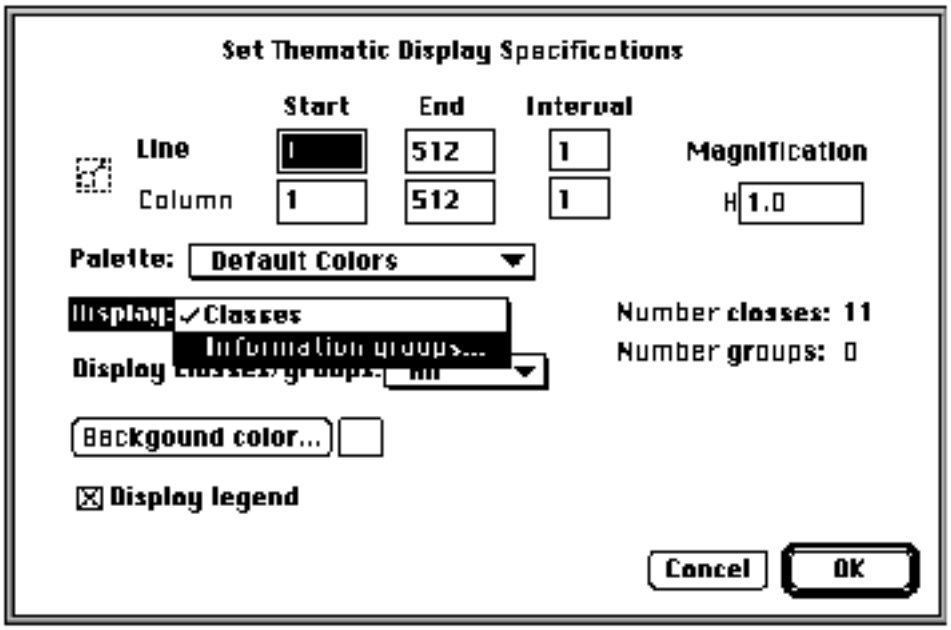
Field Verification

If there are clusters that you cannot identify "from the desk," you will have to go out into the field to determine what they are. If a formal "field trip" is not possible, in all probability someone lives near to or drives by that area and can do the identification.

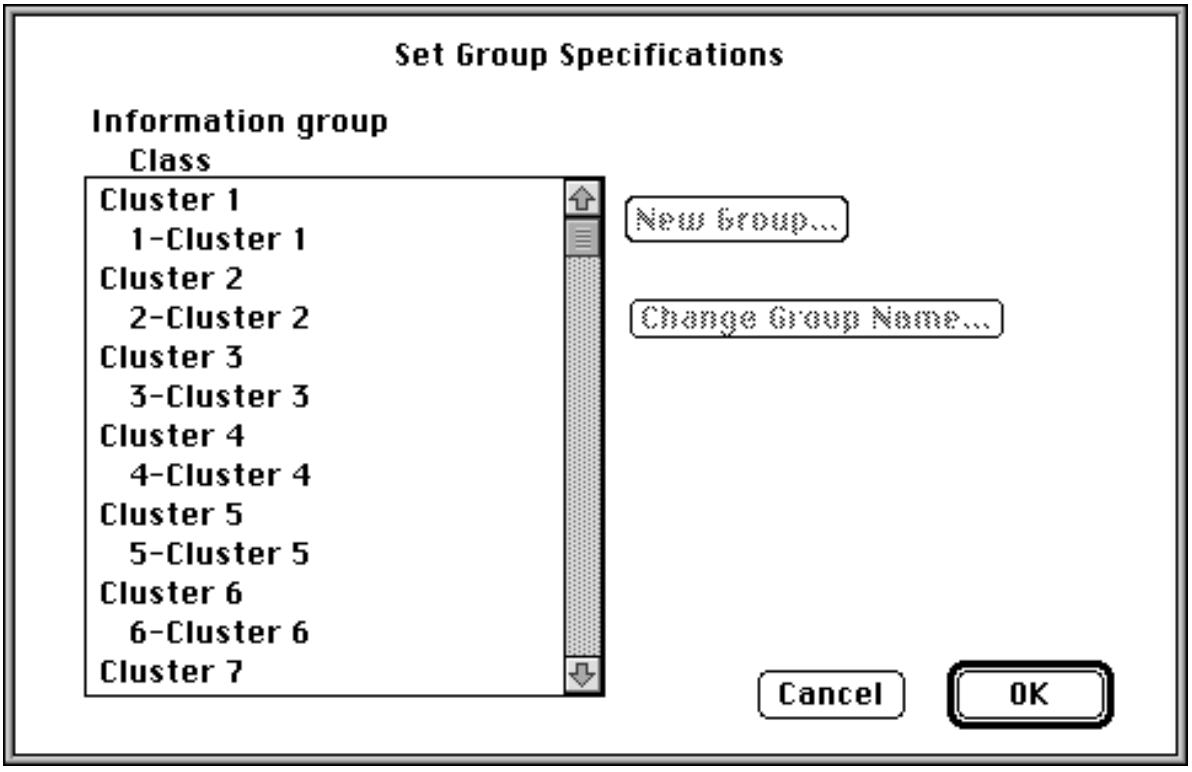
Renaming the Clusters

Your unsupervised clustering produced clusters identified only by a number, and arranged in order of decreasing brightness. You will now change the titles of these clusters to the **MUC** classification codes that you determined from your verifications.

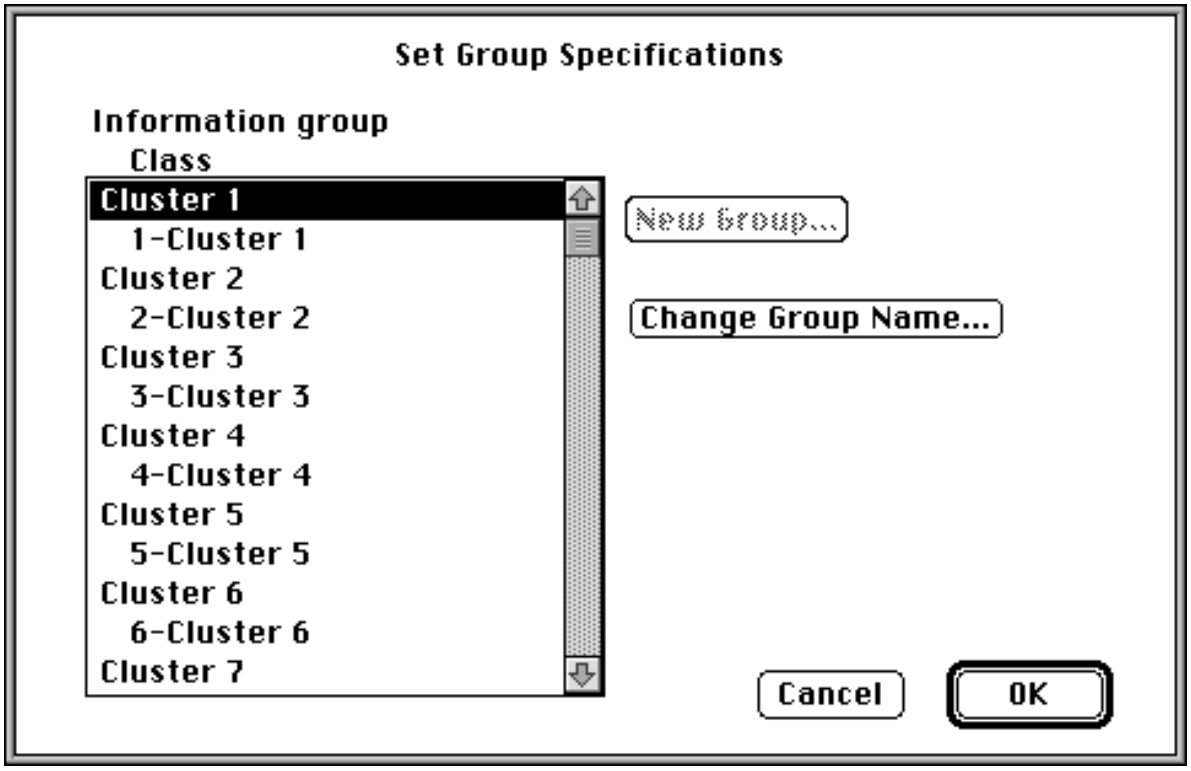
- Launch **MultiSpec**.
- From the **File** Menu, select **Open** and select your .Cluster project.
- When the **Set Thematic Display** window opens, as shown below, select **Information Groups** from the **Display** menu.



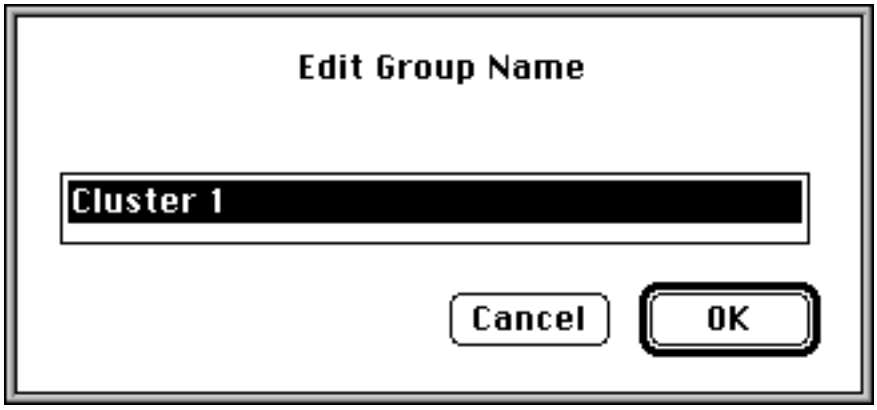
- The **Set Group Specifications** window opens, as shown below, with the "**New Group**" and "**Change Group Name**" buttons "dimmed."



- Click on "**Cluster 1**," and the dimmed "**Change Group Name**" button will darken, as shown



- Click on the **Change Group Name** button, and the **Edit Group Name** window opens, as shown below.



- Now enter, for Cluster 1, the proper **MUC Level IV designation** for the land cover represented by that cluster. Since many of the **names** of land cover types at this level are quite long, **use the MUC numerical designation** for each land cover type.
- **Repeat this process** for each of the other clusters in your map.
- You can change the colors of each of your named clusters to whatever color scheme you wish. (see page 17) When your results are sent to "Globe Central," a standard color code will be applied to it.

You have now produced a Land Use Thematic map of your 15 km x 15 km primary GLOBE Study Site.

A Note About Expectations, and a Caveat

When you proceed to the classification of your own 512 x 512 image, you will find the appearance of your clustered image probably considerably different than this demonstration. Major reasons will be:

a. This sub-set image does not contain as many land cover types as would be found in a full-sized 512 x 512 image.

b. The nature, abundance and distribution of land cover types in your image will certainly differ from those in the Beverly, Massachusetts area.

As you cluster your own image, you will find that specifying 10 clusters does not discriminate between standing bodies of water, except perhaps between fresh and salt water. In other words, lakes, ponds, rivers, etc. will probably all be clustered into the same group, unless there are significant surface properties that might change their reflectance (i.e., significant algal growth on the surface of a farm pond.)

Submitting your results:

Once you have an unsupervised classification (clustering) that seems to adequately represent your 15 x 15 GLOBE Study Site, your results will be submitted to the Land Cover and Accuracy Assessment group at the University of New Hampshire, where they will be used in ongoing studies.

Specific directions for data submission are as follows:

- Make a copy of your clustered Thematic image onto a high-density floppy disk and clearly label it with your school name, your name, and "clustered image."
- Using your favorite word processor, prepare a file with the following metadata:

  Your School Name

  Your Name

  School Address

  Date your image was acquired (If available)

  The LandSat "path and row" data, if it was supplied with your image.

  Some information about yourself, your students, and some of your experiences in doing your clustering.
- From your word processors options, save this data as a text file and place it on the same disk with your Thematic Image.
- Carefully package these disks and send them to:

Mr. Larry Ryan
Ocean Process Analysis lab
161 Morse Hall
University of New Hampshire
Durham, NH 03824 USA